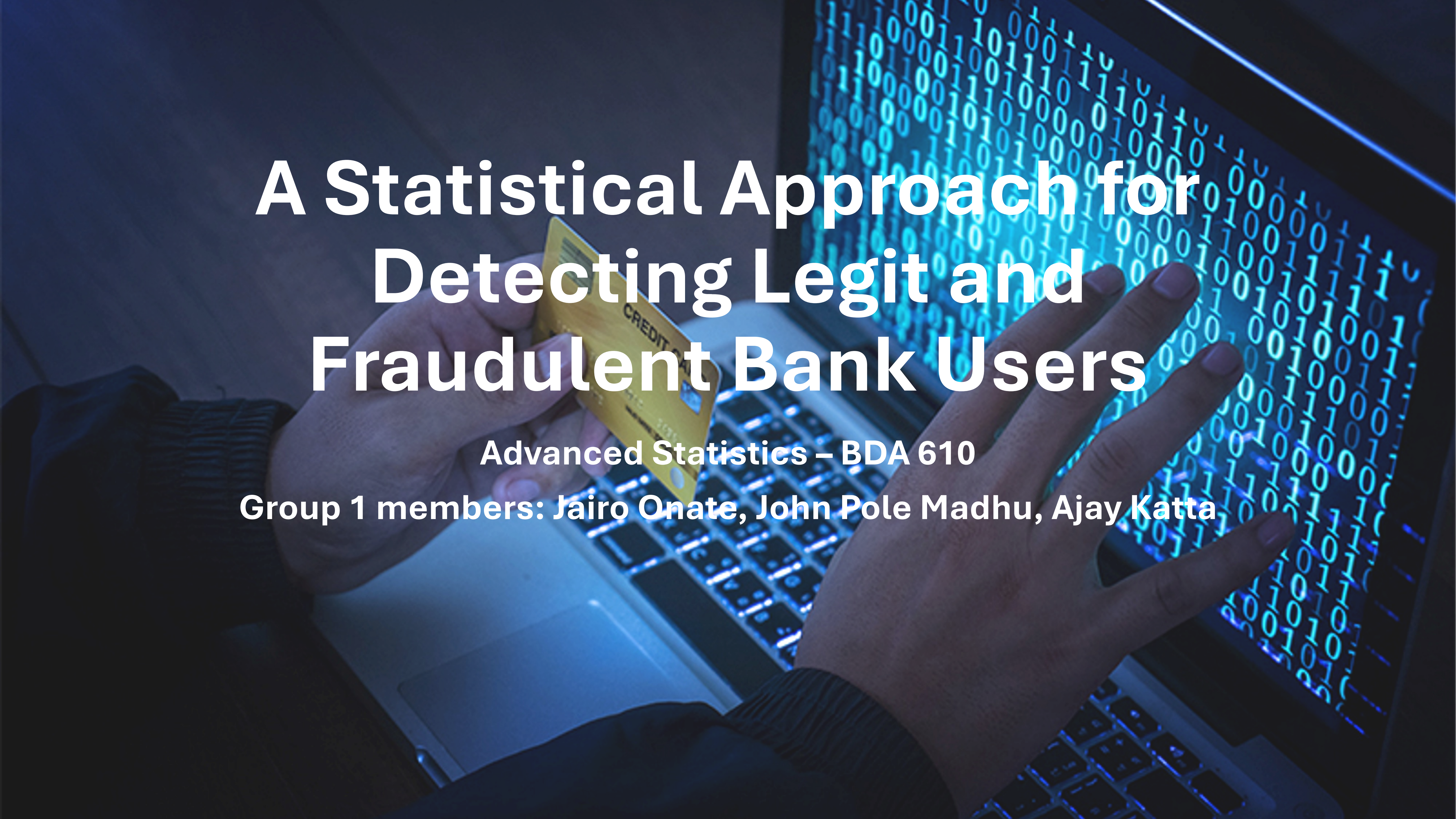


A Statistical Approach for Detecting Legit and Fraudulent Bank Users

Advanced Statistics – BDA 610

Group 1 members: Jairo Onate, John Pole Madhu, Ajay Katta

CONTENTS

1. **Introduction to Bank Fraud**

- Background
- Data
- Source

2. **Modeling**

- Data characteristics and challenges
- Feature selection and balancing

3. **Results**

- Analysis of the classification models

4. **Recommendations**

- Model selection
- Effective strategies to prevent fraud

INTRODUCTION TO BANK FRAUD

Background

- By the 1980s, advances in technology made computers accessible, leading to online banking services. Wells Fargo launched the first online platform in 1995. However, this evolution also opened doors to new types of fraud.

Problem

- In 2023, consumer losses from fraud in the U.S. totaled $10 billion (FTC). The most common fraud types were investment scams ($4.6B) and imposter scams ($2.7B).

Objective

- How can classification algorithms distinguish between legit and fraudulent bank users?

Data Source

- The dataset, titled ***Bank Account Fraud Dataset (NeurIPS 2022)***, is part of the **NeurIPS 2022 competition** and is focused on the detection of fraudulent activities associated with bank accounts

Key Factors

- Analyzing customer characteristics like annual income, email/legal name similarity, age, transfer amounts, and address history to classify a user's application as legit or fraudulent.

MODELING: DATA MANIPULATION

Dimensionality and Reduction

- Records with variables containing −1 and negative values specified by the authors as missing values were excluded from the data selection.
- Dimensionality reduction from 1,000,000 to 124,260 records with 32 variables.

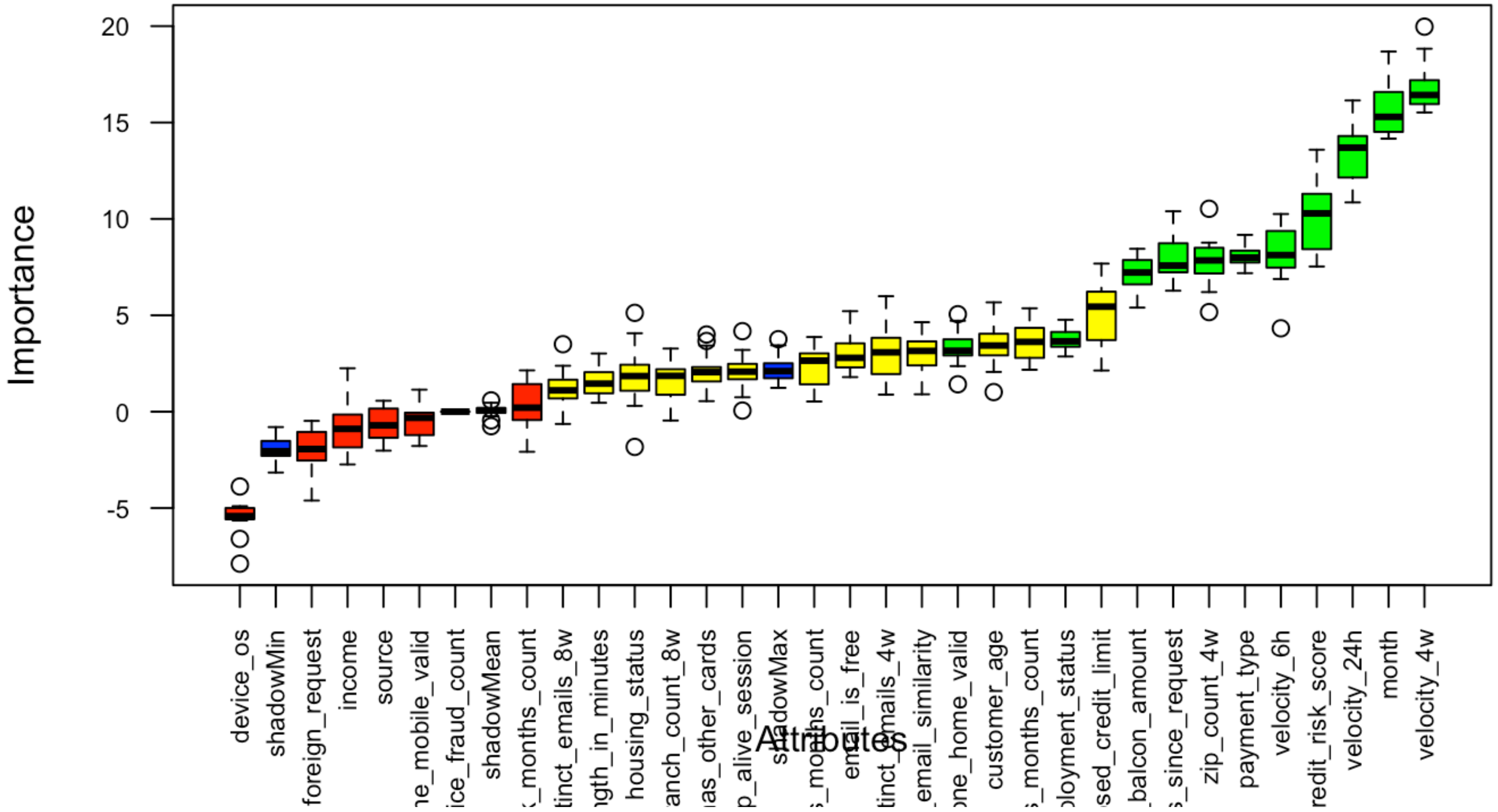
Challenges:

1. Feature Selection:

Aim: Selecting the best variables to build the model.

Tool: Boruta package which uses Random Forest to classify the importance of the attributes.

MODELING: FEATURE SELECTION



MODELING: BALANCING DATA

Challenges:

2. Balancing data:

Aim: Helping the model to prevent becoming biased towards one class.

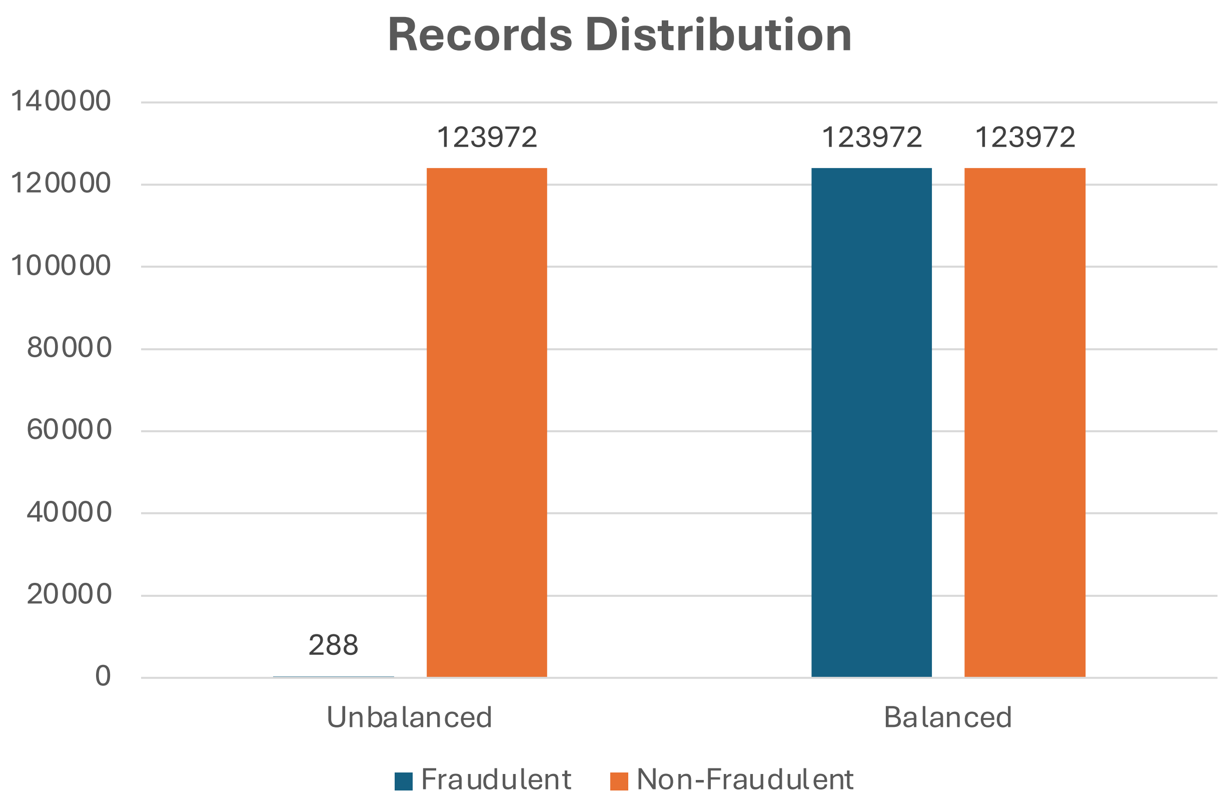
Tool: upSample method

Challenges:

1. Feature Selection:

Aim: Selecting the best variables to build the model.

Tool: Boruta package which uses Random Forest to classify the importance of the attributes.



CLASSIFICATION MODELS RESULTS

Model selected: Decision Tree

Key aspects

- **Assessment of Performance:** With 78.21% accuracy, 77.36% specificity, and 79.06% sensitivity, the Decision Tree model proved to be an excellent classifier of both fraudulent and non-fraudulent transactions.
- **Simplicity:** The concept is beneficial in real-time contexts where speed and transparency are crucial because it is computationally efficient and easy to apply.

Decision Tree - Accuracy: 78.21%		
	Reference	
Prediction	0	1
0	29456	8405
1	7801	28722

Logistic Regression - Accuracy: 70.56%		
	Reference	
Prediction	0	1
0	26714	11353
1	10543	25774

Random Forest - Accuracy: 100%		
	Reference	
Prediction	0	1
0	37257	0
1	0	37257

RECOMMENDATIONS

The Decision Tree model is a solid choice for real-world business applications where transparency is essential because it strikes a reasonable compromise between accuracy and interpretability. When a model achieves 100% accuracy, it is likely overfitting, which means it has become insensitive to noise in the training set and may not perform well on fresh data. An excessively complicated model or data leakage may be the cause of this lack of generality. Due to the inherent uncertainty in data, it is unreasonable to anticipate 100% accuracy in business; instead, it is more vital to focus on robust performance indicators like precision and recall.

Even while it falls short of Random Forest’s flawless accuracy, it offers a more approachable and realistic method of detecting fraud, which is crucial for preserving public confidence in the financial sector.